Insights into the Economy of Open Scholarship:
A look into Zenodo with Tim Smith, head of collaboration, devices and applications, CERN/IT

KE
Knowledge Exchange

DFG   CSC   Jisc   cnrs   SURF   Deff
Denmark's Electronic
Research Library

About Zenodo

The European Organisation for Nuclear Research (French: Organisation européenne pour la recherche nucléaire), known as CERN, is a European research organisation that operates the largest particle physics laboratory in the world. Established in 1954, it is based in a north western suburb of Geneva on the Franco-Swiss border, and is funded by the governments of 22 EU member states. Zenodo is the research data repository service developed and hosted by CERN. It was created in 2013 as the OpenAIRE orphan records repository. It was relaunched as Zenodo in 2015 and allows the upload of files up to 50 GB.

zenodo.org

Zenodo: Business model

Key activities

- Repository for data and other research outputs
- Option to create communities and collections
- Issues digital object identifier (DOI) for each uploaded item

Organisation type

- Non-profit
- International governmental organisation (IGO)
- Hosted at CERN

Key partners

- CERN
- Governments funding CERN
- European Commission (via OpenAIRE)

Revenue streams

- Infrastructure and staff hosted by CERN
- Project funding (OpenAIRE)
- Experimenting with crowdfunding
- Investigating a credit model with funders

IP/Copyright

- IP is always retained by the uploader
- Materials uploaded: from public domain to full copyright and all types of open licences possible
- Own code base: open source

Customers/users

- Individual researchers
- Projects
- Civil society
- Research groups
- Conference organiser
- Institutes
- National consortia

Partially based on the Business Model Canvas designed by: **Strategyzer AG (strategyzer.com)** (available under CC BY-SA 3.0)

Insights into the Economy of Open Scholarship
A look into Zenodo with Tim Smith, head of collaboration, devices and applications, CERN/IT

4

Interview with Tim Smith

CERN was founded with the idea of international collaboration to better the world. Sharing research has always been in the CERN statutes and affiliated researchers were doing this even when it wasn't called 'open' yet, such as when they invented the worldwide web and when they started preprint servers. Recently, CERN has taken a more proactive stance, sharing the software staff have created for themselves (such as Zenodo) with the entire global research community to encourage more openness in research worldwide.

"We understand that we're part of a big research ecosystem. If the system doesn't do what we want we use it nonetheless, but using our proactive approach we try to promote better solutions," says **Tim Smith**, head of collaboration, devices and applications at CERN. "We produce fundamental research that benefits the whole world: it's pure science. Open scholarship for us means that we enable others to use the same tools and techniques as we do, to follow our research or to use it in their own so that everyone benefits from our research innovations as well as our research outputs."

"Zenodo fits this ideal perfectly – we built the underlying technologies for ourselves, and then used input from other disciplines and experts to share it as a service for the entire research community."

> Open scholarship for us means that we enable others to use the same tools and techniques as we do, to follow our research or to use it in their own so that everyone benefits from our research innovations as well as our research outputs.

"Over the last decades, operating at the ever-expanding big data frontier has led us to new research workflows and to set up services drawing together computer centres across the world in a world-wide data grid," says Smith. "Opening up these datasets changes the dynamic once again. Accessible and reusable data requires even more resources than we anticipated 20 years ago, when we wrote the proposals for the projects we are still working on. With small data there's no issue, as the effort and cost is marginal. With our larger datasets, however, we have to take into account that nobody else has the tools to work with them, let alone reproduce the research.

"Overcoming these obstacles is something that we are actively looking into right now. Opening up and managing massive datasets is a learning process. What is already clear is that you need to capture relevant information as soon as you can while the research is ongoing, including the tools and metadata and intermediate datasets, otherwise reproduction of the research might become impossible afterwards."

> " Due to the scale, a lot of our research is not practically reproducible by others, so we need to duplicate and validate our research ourselves… this produces a very interesting internal dynamic. "

CERN considers itself capable of tackling most technical challenges. "The human aspect, however, is something completely different; changing existing research workflows such as publishing practices requires a lot of effort," says Smith. "At CERN, we are in a unique position with a hand in the entire research ecosystem. Due to the scale, a lot of our research is not practically reproducible by others, so we need to duplicate and validate our research ourselves by having independent apparatuses and analysis chains. This produces a very interesting internal dynamic."

"For us, it is necessary to have a 'closed' research phase wherein each experiment prepares its findings independently. Our tools therefore support the entire spectrum of research activities and outputs: from datasets that are temporarily closed, to research that has been finalised and is ready to be shared with the world."

Zenodo was created to offer other sciences the same capability to capture all research in all phases – regardless of its closed/open status. But Zenodo is also aiming for something more aspirational. Smith: "With Zenodo we can show everybody not only what can be done, but also what could and should be done when it comes to research data management. Outside of CERN all researchers can rely on this service, provided their needs are not too great. It is feasible for us, both because we want to do this for the greater good of open science and also because, as our overall infrastructure is so massive, it is a marginal activity that only requires a small addition to our resources."

Display of uploaded research article on Zenodo

At the moment, the operation of Zenodo is a peripheral activity for CERN, as the size and quantity of the datasets stored in Zenodo is easily manageable compared to its own big data. As long as Zenodo's growth remains less than the growth of CERN's own datasets, it will remain like that. Due to this unique situation, there is no urgency for a self-sustaining business model attached, as the organisation feels capable of supporting the current needs of the community without too many additional resources.

The biggest risk however seems to be that the service will become the victim of its own popularity. "If the rate of dataset growth outpaces CERN's natural data centre evolution, we'll have to incorporate this in our cash flows as it will no longer be a marginal activity. It's technically possible, but it would require a more business-oriented approach between funders, researchers, and Zenodo," says Smith. "Moreover, it needs to be taken into account that data is not static. Versioning is easy with software but applying the same techniques to mega datasets will rapidly become unmanageable if we don't keep pace with new techniques."
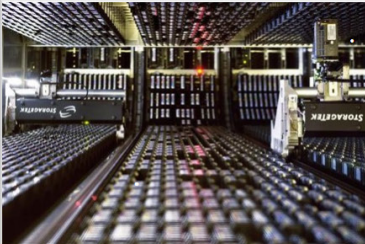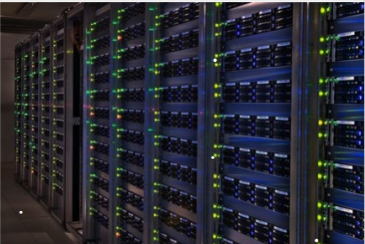
Although there's currently no formal business plan, CERN is definitely investigating potential new funding streams in order to anticipate Zenodo's sustainability.

> **Money needs to be earmarked for data services regardless, so why not invest it in a common infrastructure such as Zenodo?**

Smith: "We are investigating a credit model with major research funders. As CERN is an intergovernmentally funded organisation, one way of doing this sustainably would be to transfer funds from one national infrastructure to another via a credit system, retroactively according to usage. Money needs to be earmarked for data services regardless, so why not invest it in common infrastructures such as Zenodo? In my opinion, this would be the most efficient way to apportion existing resources to long-lasting services. But I feel funders and research-performing organisations are not quite ready for it – they see it needs many changes to cash flow processes. Our vision of shared infrastructures, common tools, and solutions requires a seachange in how funders, states, and other institutions spend their resources."

Investigating a different tack, CERN is also experimenting with using microfunding. They launched a **donation site (https://bit.ly/2PC3urO)** for Zenodo users and supporters, being completely transparent about how much it costs to run the service. "The uptake is very low though, not least because there are a lot of roadblocks in international research funding. For example, project funding cannot always be awarded for non-contract or non-commercial services. We are currently experimenting with a lot of different funding models, but our principal aim is to try and contribute to a standard shared infrastructure for managing research data. The long-term custodianship of resources that benefit the world should, in my opinion, be governed as a commons."

Big Data Founded	Cloud Connected	Human Powered
220 CHF	500 CHF	2500 CHF
1TB in Zenodo for 10 years	One day of new feature developer	One week of overall service management

Crowdfunding for Zenodo - **donation site (https://bit.ly/2PC3urO)**

At this point, Smith deliberately refuses to calculate how much working on open science costs CERN in terms of staffing: "It is a misleading number without a lot of context. Zenodo isn't run as a standalone service; it relies on the multi-level cloud services we have created in our data centre for server provisioning, service orchestration, storage, databases, and networking, which are already staffed for a very high capacity and complexity of workload."

"It also relies on the research data management programmes we have written for our own open science services. Funding via European Commission (EC) projects such as **OpenAIRE (openaire.eu)** provides the necessary extra staff to operate and support the Zenodo service on top. In addition to the grant value, we could then add a fraction of the cost of each cloud service layer and a slice of the digital library development effort. This would then represent the cost of growing a marginal service, but not the cost of creating another at CERN or elsewhere, nor the cost of running a larger scale operation. And it doesn't take account of the fact that we would maintain the same staffing level and development effort anyway!"

While CERN is not the only organisation or company offering research data management services, Smith sees a big difference from their competitors: "We are often represented as one of several similar options, but we are completely different from our colleagues at for-profit companies, the main difference being trust. Trust is a rare and hard-won currency these days and we hope that open science helps to re-establish people's trust in science in general. That's why you need organisations that are worthy of such trust, like CERN, to build shared services."

An essential element to establish this trust is, in Smith's opinion, the non-profit status of the organisation: "Not every commercial company is dishonest, it's simply that if you're driven by growth requirements you'll either monopolise the market or you'll get acquired by bigger players, that's the scheme of business."

"Therefore, I think the larger commercial operators can spawn innovation in the short term and provide valuable transient services, but in the long term they don't serve the aims of open science. Don't forget that in tech and open science even the medium term is very, very short."

"Evidence shows that no matter what guarantees the commercial companies sign up to, the moment they change leadership or get taken over they're no longer capable of fulfilling those promises.

> "For-profits can be involved in parts of the research lifecycle, but they should never be allowed to own any part of it. They can facilitate, but they cannot be the arbiter or gatekeeper."

"Commercial companies, especially startups, are very good at turning an idea into a real service. They are much better than research organisations at marketing and fundraising, and hence in service innovation, but in the consolidation phase they are often swallowed and their innovative aspect gets lost," says Smith. "That's why for-profits can usefully supply parts of the research lifecycle, but they should never be allowed to own any part of it. They can facilitate, but they cannot be the arbiter or gatekeeper, for instance, when it comes to metrics."

Contrary to competing services such as **Figshare (figshare.com)**, which in its free service only offers the most liberal **Creative Commons (https:// creativecommons.org)** licences (CC BY and CC0) as options, Zenodo allows the full set of licences including those with more restrictive clauses. Smith: "We want to enable everybody to participate in open science. This does not mean that we consider all licences as equal. But we cannot foresee what will happen in every field at any given time in the future, not even in our own field. Therefore, we think the best, the easiest, and the most persuasive service

we can give is to offer researchers what they think they need, and to gently nudge them in the direction of open science best practices in the long run."

> ❝ We think the best, the easiest and the most persuasive service we can give is to offer researchers what they think they need, and to gently nudge them in the direction of open science best practices in the long run. ❞

Zenodo is open source, based on **Invenio (https:// invenio-software.org)** software, which is a digital library framework CERN invented in the nineties. "Naturally the world has caught up on many fronts and there are now alternative components available, so we need to be flexible and agile and replace and incorporate external pieces into our framework," says Smith. "When considering externally developed

products, we always look at open source solutions first. For example, we replaced our own search engine with **Elasticsearch (elastic.co)**. We do this because we believe choosing open source is our best way of guaranteeing independence and longevity, and of increasing collaboration and promoting inclusion. But we are pragmatic at CERN, hence we operate with the best, most economical, and functional tools we can find.

"Sometimes, especially when it comes to hardware, we need to go to the biggest commercial suppliers because only they can offer the tools we need. For software, we have more flexibility and we are able to bring our principles into practice more. I use an informal decision tree that allows me to identify suitable open source tools and whether the cost of running, adapting, and maintaining them is justified. "CERN doesn't need Zenodo in order to continue its activities and Zenodo has never been intended to run at a profit. Our aim is to showcase the possibilities, to facilitate, and stimulate open science practices that were considered to be nearly impossible by most people," explains Smith.

"Ultimately Zenodo is a means to share with the world what we have learned and developed and contribute to a common and shared infrastructure for all. If somebody comes up with a better way to run such services, with equally good or better guarantees of openness, I would be happy to pass on the baton!"

" CERN doesn't need Zenodo in order to continue its activities and Zenodo has never been intended to run at a profit. Our aim is to showcase the possibilities, to facilitate, and stimulate open science practices that were considered to be nearly impossible by most people. "

References and relevant links

- ▸ CERN: **home.cern**
- ▸ Zenodo: **zenodo.org**
- ▸ Zenodo donation page: **zenodo.org/donate**
- ▸ Creative Commons licence suite: **creativecommons.org**
- ▸ Figshare: **figshare.com**
- ▸ Invenio: **invenio-software.org**
- ▸ Elasticsearch: **elastic.co**

About Tim Smith
Head of collaboration, devices and applications, CERN/IT

Tim is an open science advocate leading initiatives at CERN and in the wider science community. He drove the launch of CERN Open Data, a portal to share Large Hadron Collider (LHC) big data with the world, as well as the Higgs boson webcast, which shared its discovery live around the globe. He also instigated and nurtures Zenodo within the European Commission's OpenAIRE project as an open data service for worldwide science. Tim came to CERN at the end of the 80s, obtained a PhD in particle physics and performed research at the Large Electron-Positron Collider (LEP) for ten years. He then joined the CERN IT department to lead teams innovating in computing farm management and physics data management.